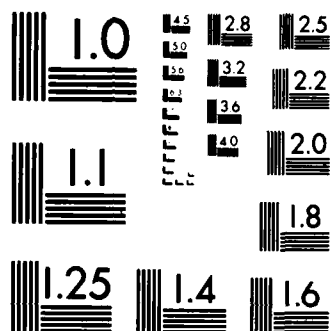MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS 1963 A

79- 0528

①

RESEARCH ON THE TECHNOLOGY OF ITEM WRITING

Gale Roid
Tom Haladyna
Teaching Research Division
Oregon State System of Higher Education

AD-A148 590

DTIC FILE COPY

Final Report

Submitted to

Defense Advanced Research Projects Agency

and

Navy Personnel Research and Development Center

June, 1979

DTIC
S **ELECTE**
DEC 1 3 1984
D
B

84 12 05 084

ABSTRACT

Research on new methods of writing test questions for use in instructional systems is summarized in this final report. Three tasks were completed: (1) experimental studies comparing several methods of transforming sentences from instructional materials into test questions, (2) a review of methods and an experimental study comparing the sentence-based questions with items written from learning objectives, and (3) the development of a Handbook on Item Writing for Criterion-Referenced Tests.

Accession For	
NTIS GRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

FOREWORD

SUMMARY

Problem

Measurement theorists have convincingly argued in recent years that there has been a lack of a scientific basis for writing achievement test items. Currently, the test specialists must provide tests that supply information for making important decisions, without a systematic technology of test-item writing. A highly developed technology of item writing is not frequently used even in the prominent method of criterion-referenced testing where an individual's performance is compared to a standard rather than to the performance of other individuals.

The most widely used methods of writing test questions, for both criterion-referenced and traditional norm-referenced tests, rely on the intuitive skills of the item writer or panels of experts who judge the merits of questions. If item writers are given learning objectives that do not precisely define the characteristics of items designed to measure the objectives, research by the authors has shown that two writers will not generate the same items or items of similar quality. A technology of item writing would help to eliminate these deficiencies in conventional methods.

Objective

The overall objective of this research was to review, describe and compare the feasibility and statistical quality of test items created by informal, objective-based and sentence-based methods of item writing. The question was posed, "To what extent do item-writer differences exist as a function of these various item-writing strategies?". There were three sub-objectives representing the three tasks of the research contract: (1) to examine statistical qualities and item-writer differences among various sentence-based methods of item writing, (2) to review and compare the statistical qualities of objective-based vs. sentence-based methods of item writing, and (3) to develop a handbook on item writing for criterion-referenced testing.

Approach

The overall approach taken in this research was first to collect and review relevant research literature and examples of implemented item-writing methods. Secondly, experimental studies were designed to systematically test differences between item types and between item writers. Thirdly, a handbook was drafted, submitted for review and subsequently revised.

The approach taken in experimental studies of item-writing methods was to select one or more units of instructional material (in all cases prose material), to define learning objectives or item-writing rules, to train three or more item writers in each method, to have each writer create several items, and to administer pretests and posttests to students who read the instructional material. All tests were composed of a balanced mixture of items of each type being contrasted. The major types of data analyses used

were (1) analyses of variance of mean differences between item difficulties (percent correct indexes) for items of each type, and (2) studies of the variability of item difficulties across item writers who were attempting to create similar items.

Findings

The major findings of the research are that item-writer differences exist and can be controlled only through quite rigorously specified item-writing rules, or field testing with subsequent revision of items to correct for these differences. An important source of difference between item writers who write multiple-choice items is the selection and wording of the wrong-answer foils. Clerical or automated methods of foil writing, as implemented in the current studies, reduced item-writer differences, but created items that tended to be easier and more susceptible to faults than items worded more freely by item writers. Some evidence from two experiments showed that cases in which item writers chose their own wording for foils resulted in items that were more sensitive to instructional effects (showing a pretest to posttest shift in difficulty) than items written by clerical methods.

A review of several methods of computerized and semi-computerized methods of item writing revealed that exemplary systems of item generation exist in scientific-technical fields such as college chemistry or military training in symbol recognition. These systems are usually computer-based and have potential for creating large banks of items. The role of the item writer becomes one of writing a computer program or set of rules rather than writing each individual item, hence, differences between item writers can be controlled. The range of item types (content or task levels) that can be created by these systems appears to be, so far, somewhat limited.

Conclusions

1. This research clearly showed that the method of writing items for criterion-referenced achievement tests can have dramatic effects on the resulting difficulty and variability between item writers of the resulting items.

2. Particularly the methods used to write foils for multiple-choice questions can have dramatic influence on the resulting difficulty and statistical quality of items.

3. Two facts that emerge from the research create somewhat of a dilemma: First, item-writing methods that give a great deal of freedom to item writers in their choice of wording result in significant differences in item difficulty between item writers, which can be an uncontrolled source of bias in criterion-referenced tests. Secondly, item-writing methods used with sentences and prose material that are clerical or computerized can result in items that are too easy, even though they control item-writer differences. This dilemma is resolved by methods which include detailed objectives or specified rules for writing items, which allow for adjustments in wording by human item writers.

4. Given that differences between item writers may exist if learning objectives or item-writing rules are used, field testing of items with student subjects becomes essential as a means of isolating and correcting for these differences.

Recommendations

1. Care should be taken that learning objectives are specific enough to correct for possible differences between item writers who interpret the requirements for each objective.

2. When using multiple-choice items, documentation of the methods used to select the wrong-answer foils for each item should be developed during Phase II, Step 2, of Instructional Systems Development (ISD).

3. Field testing and empirical item analysis, as well as review by subject-matter experts, should be regularly used to identify and isolate possible item-writer differences in the construction of items for criterion-referenced tests.

4. It is recommended that a needs assessment be conducted of areas in military training where prose instructional materials and reading-comprehension tests are used. Where such a need is found, further research, development, and application of the sentence-based item-writing methods created by the current research should be explored.

CONTENTS

INTRODUCTION

Three major tasks were completed on this research contract. Task 1 was to conduct and report a study comparing different procedures for writing multiple-choice questions from sentences in instructional materials. Task 2 was to conduct and report a study comparing sentence-based test questions and objective-based test questions. Task 3 was to develop and revise a handbook on item writing for criterion-referenced tests. This final report summarizes the results of each task and the reports furnished as deliverables on the contract. In addition, a number of extra reports were produced from the research and these are described along with the contract deliverables. Deliverables and products of the research are listed in Table 1, and each is described in the remainder of this report.

Table 1

Contract Deliverables and Products

1. Interim Technical Report, Task 1

2. Technical Report, Task 1

3. Technical Paper, Task 1

4. Additional Reports, Task 1

5. Additional Experiment, Task 1

6. Technical Report, Task 2

7. Technical Paper, Task 2

8. Published Book Chapter, Task 2

9. Handbook on Item Writing (First Draft), Task 3

10. Handbook on Item Writing, Task 3

11. Book To Be Published

TASK 1: STUDIES OF SENTENCE-BASED ITEMS

Interim Technical Report, Task 1

A pilot study was conducted to compare several procedures for transforming sentences from instructional materials into test questions. This Interim Technical Report was filed with the Navy Personnel Research and Development Center on August 31, 1977. It was subsequently published in June, 1978, as NPRDC Technical Report 78-23 entitled, "Algorithms for Developing Test Questions from Sentences in Instructional Materials." In this study, a computer-based algorithm was used to analyze prose subject matter and to identify high-information words. These words were keywords in sentences, and were either nouns or adjectives. The recommendations of this study were that infrequently occurring nouns and adjectives and frequently occurring adjectives should be used to select sentences from prose passages for transformation into questions that measure reading comprehension. Frequently occurring nouns should not be used for questions, particularly when they occur in general introductory sentences. Also, it was recommended that methods of algorithmically generating the wrong-answer foils for multiple-choice questions should be further refined and applied in a variety of subject-matter areas.

Technical Report, Task 1

Technical Report #1 was filed in February, 1978, entitled, "A Comparison of Methods for Transforming Sentences into Test Questions for Instructional Materials." This study examined the idea that methods of writing test questions, particularly for criterion-referenced tests, should be based on operationally defined rules. This study was designed to examine and further refine a method for objectively generating multiple-choice questions for prose instructional materials. Important sentences were selected from a

prose passage in a science text and these sentences were transformed into questions. Several variations of sentence transformation rules were used to create tests given to 273 college and high school students before and after they read the passage. Item difficulties (percent correct) for each type of item formed the basic data of the study. The study concluded that the method of selecting the "question word" (a noun or adjective) in the sentence has a crucial role in determining the pattern of pretest and posttest item difficulties of the resulting question. Also, the methods of item writing used in the study were found to be feasible and to be relatively free from the item-writer differences that typically are found in traditional item-writing methods.

Technical Paper, Task 1

An adapted version of Technical Report, Task 1, was presented at the meetings of the American Educational Research Association in Toronto, March, 1978. This paper was entitled, "A Comparison of Several Multiple-Choice Linguistic-Based Item-Writing Algorithms," authored by Gale Roid and Tom Haladyna. The paper was part of a symposium organized by Tom Haladyna which included three other papers and discussions by other researchers in the field. Several of the consultants to this contract were participants in the symposium. The symposium had a large audience and, as a result, more than 150 copies of this paper have been distributed to educational researchers and measurement experts across the country. A slightly revised version of this paper was submitted for publication to the Journal of Reading Behavior, and we are awaiting word from the editorial board.

Additional Reports, Task 1

A preliminary version of Technical Report #1 was presented at the conference sponsored by Defense Advanced Research Projects Agency at Spring Hill, Minnesota, August, 1977. The paper was entitled, "A Linguistic Basis for Developing Tests," as part of the seminar entitled, "Innovation in Instructional Systems Development." This conference was attended by other DARPA contractors, military and university training experts, and was an excellent forum for disseminating the research on the present contract.

A second brief report on Task 1 was written by the Project Director and subsequently published in the Computer Assisted Test Construction Digest (CATC Digest). The article entitled, "Computer Analysis for Question Writing," was published in Volume 2, Number 4, 1978, p. 4. The CATC Digest is published by the Educational Testing Service, Princeton, New Jersey. A copy of this article is attached as Appendix A.

Additional Experiment, Task 1

In order to further refine several of the sentence-based item-writing methods, a second experiment was conducted and reported as Technical Report #3 entitled, "Item Writing for Domain-Based Tests of Prose Learning." This additional technical report was filed in November, 1978. This study examined several methods of writing test questions that may solve the problems of optimizing the match between teaching and testing and controlling item-writer differences. Specifically, rules for writing test items from prose instructional material were given to four item writers who created multiple-choice test items. Differences among item writers and among a variety of item-writing rules were examined. Tests were given to 423 students before and after reading two prose passages. Methods of transforming sentences from prose

material into test questions were found to control the typical variance of item difficulty that is observed between item writers. The information density of the prose passage, the method of writing foils (wrong-answer alternatives), the part of speech of the keywords in sentences transformed into questions, and verbatim vs. paraphrase use of sentences had important influence on the statistical characteristics of items. It is recommended that a needs assessment be conducted to identify areas in military training where the methods of the present study could be implemented, and to field test these.

A technical paper for this additional experiment was presented at the meetings of the American Educational Research Association in San Francisco, April, 1979, under the title, "Item Writing for Domain-Referenced Tests of Prose Learning," authored by Roid and Haladyna.

TASK 2: COMPARISON OF OBJECTIVE-BASED AND SENTENCE-BASED ITEM-WRITING METHODS

Technical Report, Task 2

Task 2 of the present research project was intended to be a comparison of the best sentence-based methods developed under Task 1 with objective-based methods of writing test items. This Technical Report #4, entitled, "A Comparative Study of Informal, Objective-Based and Linguistic-Based Item-Writing Methods," authored by Roid, Haladyna and Shaughnessy, was submitted in December, 1978. This study compared the statistical qualities of items written by six item writers who used a variety of informal and objective methods for constructing questions. The six item writers developed pretests and posttests for a unit from a children's wildlife magazine. Item responses of 364 elementary school students who were given instruction on the unit

5

were tabulated and item difficulties (percent correct responses) were used as the basic data of the study. The study clearly showed that the method of writing test items, and particularly the method by which foils (wrong-answer alternatives) were created, had significant effects on the pattern of item difficulties of the resulting items. Informal methods of item writing, in which item writers have maximal freedom in choice of wording, resulted in large differences between experienced item writers and teachers. A clerical method of writing foils was shown to produce items that were too easy. The study indicates the importance of field testing and analyzing test items to identify possible differences between item writers that may cause an uncontrolled source of bias in criterion-referenced tests.

Technical Paper, Task 2

In order to compare objective-based, sentence-based and other methods of item writing, a comprehensive review paper was prepared for Task 2. The first version of this review paper was presented at the annual meeting of the Military Testing Association, Oklahoma City, October, 1978, under the title, "A Review of Item-Writing Methods for Criterion-Referenced Tests in the Cognitive Domain." Dr. Haladyna presented the paper to a large and receptive audience at this meeting. This review paper was then revised and improved and submitted for publication in the Proceedings of the Military Testing Association. The Proceedings were published by the U. S. Coast Guard Institute in the spring of 1979 and the review paper appears in Volume 2, pp. 1,035-1,066, under the title, "The Emergence of an Item-Writing Technology." This paper provides a review of the emerging technology of test-item writing for criterion-referenced tests. A continuum of item-writing methods is proposed, ranging from informal-subjective methods to automated-objective methods. Examples of techniques include objective-based item

writing, amplified objectives, item forms, facet design, domain-referenced

concept testing and computerized techniques. Data from studies of item-

writing techniques are also reviewed. Recommendations for futher research

and for applications to criterion-referenced testing are presented.

A version of this review paper in technical report format was submitted

as Technical Report #2 to the Navy Personnel Research and Development Center

and DARPA in November, 1978, entitled, "A Review of Item-Writing Methods for

Criterion-Referenced Tests," authored by Roid and Haladyna. This paper has

been submitted for publication in the journal, Review of Educational Research.

Published Book Chapter, Task 2

As part of the continuing effort to disseminate results of this research

contract, an invited book chapter was written and subsequently published

entitled, "The Technology of Test-Item Writing." This chapter takes a

similar approach to the Technical Paper, Task 2, in reviewing the various

methods of test-item writing. The book of which this chapter is a part is

ready for release by the publishers, Academic Press, in the summer of 1979.

The reference for this chapter is as follows:

> Roid, G. The technology of test-item writing. In
> Harold F. O'Neil, Jr. (Ed.), Procedures for
> instructional systems development. New York:
> Academic Press, 1979, pp. 67-94.

TASK 3: HANDBOOK ON ITEM WRITING FOR CRITERION-REFERENCED TESTS

Handbook on Item Writing, Task 3

Task 3 of this research contract was to develop, evaluate and revise a

handbook on item writing for criterion-referenced tests. The project staff

worked closely with the Contracting Officer's Technical Representative,

Dr. Pat-Anthony Federico, of the Navy Personnel Research and Development

Center in the design of this handbook on item writing. A first draft of the handbook was submitted on January 15, 1979, and consisted of 280 pages of manuscript. The objective of this handbook was to train instructors and test developers in the military in writing high quality, criterion-referenced test items. Seventeen chapters were included in this draft handbook.

Table 2

Chapters in First Draft of Item-Writing Handbook

Chapter 1: Why Read This Handbook and How To Use It

Chapter 2: Fundamental Concepts of Testing in Systematic Instruction

Chapter 3: A Framework for Criterion-Referenced Testing in the Cognitive Domain

Chapter 4: Selected-Response Test Items

Chapter 5: Writing Constructed-Response Test Questions

Chapter 6: Writing Items from Prose Learning: Making Sentences into Questions

Chapter 7: Writing Objective-Quantitative Items

Chapter 8: Test Items for Concepts

Chapter 9: Measuring Higher-Level Thinking

Chapter 10: Measuring Skills: Performance or Product

Chapter 11: Constructing and Properly Using Rating Scales

Chapter 12: Measuring Skills Through Observation

Chapter 13: Evaluating Skills Through the Use of Checklists

Chapter 14: Empirical Review of Knowledge Items

Chapter 15: Empirical Item Review for Skills

Chapter 16: The Logical Review of Criterion-Referenced Test Items

Chapter 17: The Technology of Item Writing: Summary and Conclusions

The COTR subsequently conducted a review of this manual with several readers who provided comments on the content of the handbook and its usefulness in military training. The result of this review was a two-fold recommendation:

1. That the handbook be redesigned to be more brief and concise, and

2. That the handbook be made to match the methods of the Instructional Quality Inventory (IQI) published by the Navy Personnel Research and Development Center. The IQI is a comprehensive method for evaluating the consistency and adequacy of objectives, test items and instructional materials, and is used heavily by Navy training personnel. For this reason, a handbook that was coordinated with the IQI could potentially be widely implemented.

Therefore, a revised handbook was designed and was submitted on June 30, 1979, containing nine chapters and approximately 58 pages, as shown in Table 3.

Table 3

Chapters in Final Draft of Item-Writing Handbook

Chapter 1: Introduction to Item Writing

Chapter 2: Recognition Test Questions

Chapter 3: Recall Test Questions

Chapter 4: Measuring Performances and Products

Chapter 5: Rating Scales

Chapter 6: Measuring Performances or Products Through Observation

Chapter 7: Checklists

Chapter 8: Logical Item Review

Chapter 9: Field Testing of Items

Chapters 1 and 8 of this handbook reference the Instructional Quality Inventory (IQI) and demonstrate how the reader can coordinate the methods in this handbook with the IQI procedures.

Additional Publication: Book for Academic Press

As a result of this research contract, Drs. Roid and Haladyna were invited by Academic Press to write a book intended for educational researchers on the technology of test-item writing. A manuscript for the book, to be entitled "A Technology for Test-Item Writing," is due December 31, 1979. At the present time it is the intention of the authors to revise and improve the larger first draft of the item-writing handbook for adaptation to a book publication. This will require removal of exercises, Chapter tests and objectives and insertion of additional references to research publications.

ADAPTIONS OF THE ORIGINAL WORK PLAN

The major adaption of the original work plan and proposal was the obtaining of two no-cost extensions, the first from August to December, 1978, and the second from December, 1978, to June 30, 1979. These extensions were granted to allow for the extensive review and planning that went into the handbook on item writing, which took a different form and concept, in order to fit the needs of the military, than was originally proposed. The original proposal was for a more scholarly research-type handbook. However, it became clear through meetings with the COTR that the real need for a handbook was two-fold: (1) to be used in military training programs by instructors who did not have measurement background, and (2) that the handbook be coordinated with the Instructional Quality Inventory. Another positive adaption of the original work plan was the addition of extra papers and

reports. A large number of publishable papers, a book chapter, and a book
are outgrowths of this research which were not originally anticipated.
Therefore, a great deal more dissemination has been possible because of
the support from this research contract than was originally expected. Other
minor changes in the experimental procedures were necessitated during the
conduct of the experiments. It was found that sample sizes and numbers of
test items needed to be increased from the original proposal to provide
more statistical precision. This necessitated a reduction in the number of
types of samples used (e.g., dental school tests and tests of quantitative
subject matter were eliminated).

CONCLUSIONS

A good deal of dissemination and a larger number of written reports than originally anticipated were produced through the support of this research contract. Within the next two years, it is anticipated that additional published articles from this research will appear in print.

Three research experiments were completed, and they showed that the method of item writing used by a test developer can have dramatic effects on the difficulty and other characteristics of the resulting items. Particularly, the method of writing the wrong-answer foils for multiple-choice questions can have strong effects on the characteristics of items. In cases where prose instructional materials are used, and students are given tests of reading comprehension, the current research provides several methods for generating sentence-based items for criterion-referenced tests of prose learning. These algorithmic or clerical methods of item writing appear to control differences between item writers. However, some evidence indicates that this control of differences comes at the expense of creating items that are too easy. Because this suggests that further refinements are necessary in sentence-based methods before their widespread use, the Handbook on Item Writing produced by the research contract was revised to emphasize objective-based methods of item writing. This is done with the caution that evidence from the experiments shows that item-writer differences will be present in objective-based methods. Therefore, it is concluded that the following controls on item writing should be used: (1) detailed specifications within the learning objectives, (2) use of the Instructional Quality Inventory to evaluate items, and (3) subsequent field testing and revision of items.

RECOMMENDATIONS

1. Care should be taken that learning objectives are specific enough to correct for possible differences between item writers who interpret the requirements for each objective.

2. When using multiple-choice items, documentation of the methods used to select the wrong-answer foils for each item should be developed during Phase II, Step 2, of Instructional Systems Development (ISD).

3. Field testing and empirical item analysis, as well as review by subject-matter experts, should be regularly used to identify and isolate possible item-writer differences in the construction of items for criterion-referenced tests.

4. It is recommended that a needs assessment be conducted of areas in military training where prose instructional materials and reading-comprehension tests are used. Where such a need is found, further research, development, and application of the sentence-based, item-writing methods created by the current research should be explored.

APPENDIX A

CATC Digest Article

Computer Analysis for Question Writing

by Gale Roid

How do you identify the important elements that a student should remember from a text? One approach to this problem has been proposed by Patrick Finn of the State University of New York at Buffalo, who uses a computer analysis of words in a prose passage. All the words in a passage are keypunched for input to a computer program that performs two major tasks: counts the number of times that each word appears in the passage, and identifies the standard frequency index (SFI) of each word.

The SFI of each word is a numerical estimate of how often the word occurs in a large corpus of American English, described in J.B. Carroll, P. Davies, and B. Richman's *Word Frequency Book* (1971). A computer tape containing the SFI of more than five million words is used to identify the SFI of each word in a passage. The tape is a computerized version of the Carroll, Davies, and Richman book.

SFI's range from 88.6 for "the" (meaning that the average American student is likely to encounter this word once in every ten words of school book reading) to 02.5 for "incarnation" (the average student is likely to encounter this word less than once out of every million words).

The goal of this kind of analysis is to identify "high information" words—those that are relatively rare in American English and occur only a single time in a given passage. The sentences in which these high information words occur can then become candidates for transformation into questions that tap important information in the passage. High information words are those which might be difficult for students to remember if they were not tested on these elements.

Finn has done considerable research in developing this method of identifying high information words and has a linguistic theory explaining the method. Once an important word has been identified, the sentence in which it occurs can be transformed into a question, using the methods of J.R. Bormuth (*On the Theory of Achievement Test Items,* 1970) and Finn ("A Question Writing Algorithm," *Journal of Reading Behavior,* 1975).

In Oregon, we have been experimenting with several methods based on Finn's techniques, and our experience shows that not all parts of speech are equally good candidates for developing questions, even though they may be high information words. Verbs and adverbs in particular are difficult words to remove from a sentence that is transformed into a question. After considerable attempts to produce questions from verbs, Finn and I have concluded that the most promising parts of speech are adjectives and nouns. Some recent research available in a technical report (Roid and Finn, 1977) has shown the feasibility of this method for analyzing prose and writing test questions. ■

Dr. Gale Roid is Associate Research Professor, Teaching Research Division, Oregon State System of Higher Education, Monmouth, OR 97361.

CATC through ERIC

The Educational Resources Information Center (ERIC) clearinghouses form a national information system that provides ready access to results of educational programs and research and development efforts.

ERIC/TM, located at ETS, is responsible for collecting, organizing, and disseminating current information on testing, measurement, evaluation, learning theory, and human development. Over 130,000 documents on microfiche comprise the complete ERIC collection which can be accessed through a computer retrieval system. Reports are available in microfiche or text.

To request a literature search or submit a document, write: ERIC/TM, ETS, Princeton, NJ 08541.

END

FILMED

1-85

DTIC